Hey! What's New? 2024-102

Research Uncovers 'Critical' Knowledge Gaps in AI Governance

According to a just issued press release from MIT CSAIL, "as organizations rush to implement artificial intelligence (AI), a new analysis of AI-related risks finds significant gaps in our understanding, highlighting an urgent need for a more comprehensive approach."

The research notes that "adoption of AI is rapidly increasing; census data shows a significant (47%) rise in AI usage within US industries, jumping from 3.7% to 5.45% between September 2023 and February 2024. However, a comprehensive review from researchers at MIT CSAIL and MIT FutureTech has uncovered critical gaps in existing AI risk frameworks. Their analysis reveals that even the most thorough individual framework overlooks approximately 30% of the risks identified across all reviewed frameworks."

To help address this, they collaborated with colleagues from the University of Queensland, Future of Life Institute, KU Leuven and Harmony Intelligence to release the first-ever *AI Risk Repository*: a comprehensive and accessible living database of 700+ risks posed by AI that will be expanded and updated to ensure that it remains current and relevant.

"Since the AI risk literature is scattered across peer-reviewed journals, preprints and industry reports, and quite varied, I worry that decision-makers may unwittingly consult incomplete overviews, miss important concerns, and develop collective blind spots," says Dr. Peter Slattery, an incoming postdoc at the MIT FutureTech Lab and current project lead.

After searching several academic databases, engaging experts, and retrieving more than 17,000 records, the researchers identified 43 existing AI risk classification frameworks. From these, they extracted more than 700 risks. They then used approaches that they developed from two existing frameworks to categorize each risk by cause (e.g., when or why it occurs), risk domain (e.g., "Misinformation"), and risk subdomain (e.g., "False or misleading information").

Examples of risks identified include "Unfair discrimination and misrepresentation," "Fraud, scams and targeted manipulation" and "Overreliance and unsafe use." More of the risks analyzed were attributed to AI systems (51%) than humans (34%) and presented as emerging after AI was deployed (65%) rather than during its development (10%). The most frequently addressed risk domains included "AI system safety, failures and limitations" (76% of documents); "Socioeconomic and environmental harms" (73%); "Discrimination and toxicity" (71%); "Privacy and security" (68%); and "Malicious actors and misuse" (68%). In contrast, "Human-computer interaction" (41%) and "Misinformation" (44%) received comparatively less attention.

The work addresses the urgent need to help decision-makers in government, research and industry understand and prioritize the risks associated with AI and work together to address them. "Many AI governance initiatives are emerging across the world focused on addressing key risks from AI," says collaborator Risto Uuk, EU Research Lead at the Future of Life Institute. "These institutions need a more comprehensive and complete understanding of the risk landscape."

"There's a significant need for a comprehensive database of risks from advanced AI which safety evaluators like Harmony Intelligence can use to identify and catch risks systematically," argues collaborator Soroush Pour, CEO & Co-founder of AI safety evaluations and red teaming company Harmony Intelligence. "Otherwise, it's unclear what risks we should be looking for, or what tests need to be done. It becomes much more likely that we miss something by simply not being aware of it".

"The AI Risk Repository is, to our knowledge, the first attempt to rigorously curate, analyze, and extract AI risk frameworks into a publicly accessible, comprehensive, extensible, and categorized risk database. It is part of a larger effort to understand how we are responding to AI risks and to identify if there are gaps in our current approaches," says Dr. Neil Thompson, head of the MIT FutureTech Lab and one of the lead researchers on the project. "We are starting with a comprehensive checklist, to help us understand the breadth of potential risks. We plan to use this to identify shortcomings in organizational responses. For instance, if everyone focuses on one type of risk while overlooking others of similar importance, that's something we should notice and address."

For much more, see Global AI adoption is outpacing risk understanding, warns MIT CSAIL | MIT CSAIL.